Limits of logic-based inherent safety of social robots

Extended abstract

Bentzen, Martin Mose

Publication date:
2014

Extended abstract accepted for the fPET conference, Virginia Tech, USA, May 27-29, 2014

Limits of logic-based inherent safety of social robots

Martin Mose Bentzen, DTU Management Engineering, Technical University of Denmark

Social robots can reason and act while taking into account social and cultural structures, for instance by complying with social or ethical norms or values. Voices within industry and governments, as well as within academia, are predicting that in the future social robots will enter very different places of human life, e.g. hospitals, homes for the elderly, battle fields, museums, and so on. As social robots are likely to become more common and advanced and thus likely to interact with human beings in increasingly complex situations, considering risks and ensuring safety in such situations will become very important. This paper investigates various aspects of the safety of social robots from a conceptual perspective. First, a distinction between low-level operational risks and high-level social risks will be made. An example of a low-level risk is a robot injuring a human being because of a software glitch or because of failure in human-robot communication. Examples of high-level social risks are a child suffering a mental disorder from being raised solely by a robot or an elderly person suffering from intense loneliness because of solely having contact with social robots. It is shown that there can be bridging problems between the low-level functionality of social robots and their high-level functionality, i.e. even if social robots function satisfactorily according to low-level task analyses, they may nonetheless pose high-level social risks to their users.

I then consider one very important strategy with regard to robot safety. The safety strategy considered will be called "logic-based inherent safety of social robots". The idea, as suggested first in science fiction by Asimov but in recent times also by robot theorists such as Bringsjord, Arkoudas, Bello, Lokhorst and others, is that robots should be logically guaranteed to act in a socially acceptable or ethical way. Bringsjord and Taylor have defined ethically correct robots, as robots which satisfy the following three core desiderata.

D1 Robots take only permissible actions. D2 All relevant actions that are obligatory are actually performed by them, subject to ties and conflicts among relevant actions. D3 All permissible (or obligatory or forbidden) actions can be proved by the robot to be permissible (or obligatory or forbidden) and all such proofs can be explained in ordinary English. (Bringsjord and Taylor (2012))

As attractive as these desiderata may seem to the designers and users of social robots, it is shown by a Gödelian argument that, given what seems a reasonable interpretation of the above desiderata, it is not logically possible for a robot to fulfil all three desiderata at the same time. If we require a logic-based robot (all current robots are) to be ethically correct (and thus only perform actions it is obliged to perform), it cannot also prove itself to be ethically complete (and thus prove every obligation it fulfils). There will be actions which are ethically obligatory from a meta-logical point of view, but which cannot be proven by the robot to be so. An example is a robot, R, logically equipped to express a sentence, such as "R ought to prove sentence s". Here "s" could be a sentence with an ethical content. But then we can construct the sentence g: "R ought not to prove sentence g". If R is correct it cannot prove g, because by proving it, it will have violated the obligation expressed in the sentence. On the other hand, R fulfils the obligation expressed in g by simply not proving it. Thus there is an obligation fulfilled by the robot which cannot be proven to be obligatory by the robot. After this, I show some more problems with logic-based inherent safety. One problem is related to the myth of logic as a more "literal language" which gets to the real meaning of words (it is not). This is made as a general point and also exemplified by showing a situation where the relationship between deontic logic and natural language could lead to a serious failure in Human-robot communication. I

conclude with cautious optimism that deontic logic will have a big role to play in the design of safe ethical robots, but only if the following criteria are met: 1) The use of robots is restricted to low-level functions. 2) Logics closer to natural language than most current deontic logics are devised. 3) Even such a logic only plays a limited part of the overall safety of the robot, itself subject e.g. to negative feedback safety mechanisms and independent safety barriers.

Bringsjord, S. and Taylor, J. (2012). The divine-command approach to robot ethics, in P. Lina, K. Abney and G. A. Bekey (eds), Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, pp. 85-108.